Clasificación Binaria

Estudiantes de Portugués

Sergio Del Castillo Baranda

4/10/2020

Carga de los datos y librerías

```
students.csv <- file.path(getwd(), 'student-por.csv')
STUDENTS <- read.csv2(file = students.csv, header = TRUE, sep = ';')

summary(STUDENTS)
```

```
##  school   sex          age        address famsize   Pstatus
##  GP:423   F:383   Min.   :15.00   R:197   GT3:457   A: 80
##  MS:226   M:266   1st Qu.:16.00   U:452   LE3:192   T:569
##                   Median :17.00
##                   Mean   :16.74
##                   3rd Qu.:18.00
##                   Max.   :22.00
##       Medu            Fedu             Mjob           Fjob
##  Min.   :0.000   Min.   :0.000   at_home :135   at_home : 42
##  1st Qu.:2.000   1st Qu.:1.000   health  : 48   health  : 23
##  Median :2.000   Median :2.000   other   :258   other   :367
##  Mean   :2.515   Mean   :2.307   services:136   services:181
##  3rd Qu.:4.000   3rd Qu.:3.000   teacher : 72   teacher : 36
##  Max.   :4.000   Max.   :4.000
##         reason      guardian     traveltime      studytime
##  course    :285   father:153   Min.   :1.000   Min.   :1.000
##  home      :149   mother:455   1st Qu.:1.000   1st Qu.:1.000
##  other     : 72   other : 41   Median :1.000   Median :2.000
##  reputation:143                Mean   :1.569   Mean   :1.931
##                                3rd Qu.:2.000   3rd Qu.:2.000
##                                Max.   :4.000   Max.   :4.000
##     failures      schoolsup famsup     paid     activities nursery
##  Min.   :0.0000   no :581   no :251   no :610   no :334    no :128
##  1st Qu.:0.0000   yes: 68   yes:398   yes: 39   yes:315    yes:521
##  Median :0.0000
##  Mean   :0.2219
##  3rd Qu.:0.0000
##  Max.   :3.0000
##  higher    internet  romantic      famrel         freetime
##  no : 69   no :151   no :410   Min.   :1.000   Min.   :1.00
##  yes:580   yes:498   yes:239   1st Qu.:4.000   1st Qu.:3.00
##                                Median :4.000   Median :3.00
##                                Mean   :3.931   Mean   :3.18
```

```
##                                3rd Qu.:5.000   3rd Qu.:4.00
##                                Max.   :5.000   Max.   :5.00
##      goout            Dalc            Walc          health
##  Min.   :1.000   Min.   :1.000   Min.   :1.00   Min.   :1.000
##  1st Qu.:2.000   1st Qu.:1.000   1st Qu.:1.00   1st Qu.:2.000
##  Median :3.000   Median :1.000   Median :2.00   Median :4.000
##  Mean   :3.185   Mean   :1.502   Mean   :2.28   Mean   :3.536
##  3rd Qu.:4.000   3rd Qu.:2.000   3rd Qu.:3.00   3rd Qu.:5.000
##  Max.   :5.000   Max.   :5.000   Max.   :5.00   Max.   :5.000
##     absences            G1              G2              G3
##  Min.   : 0.000   Min.   : 0.0   Min.   : 0.00   Min.   : 0.00
##  1st Qu.: 0.000   1st Qu.:10.0   1st Qu.:10.00   1st Qu.:10.00
##  Median : 2.000   Median :11.0   Median :11.00   Median :12.00
##  Mean   : 3.659   Mean   :11.4   Mean   :11.57   Mean   :11.91
##  3rd Qu.: 6.000   3rd Qu.:13.0   3rd Qu.:13.00   3rd Qu.:14.00
##  Max.   :32.000   Max.   :19.0   Max.   :19.00   Max.   :19.00
```

a) Se deben realizar pruebas suficientes para obtener una buena selección de variables, obteniendo uno o varios conjuntos de variables tentativos

En primer lugar voy a separar en dos variables el conjunto de columnas del dataset que son categóricas y las que son continuas. A continuación tendremos que realizar la estandarización únicamente de las continuas.

Em este apartado seleccionamos la variable objetivo

```
continuas <- dput(names(select_if(STUDENTS[,-21], is.integer)))
categoricas <- dput(names(select_if(STUDENTS[,-21], is.factor)))

vardep <- "higher"

higher <- STUDENTS[,vardep]

cat("Variables continuas: ",continuas, "\nVariables categoricas: ",categoricas,"\nNuestra variable obje
```

Eliminar las observaciones con missing en alguna variable

```
STUDENTS<-na.omit(STUDENTS,(!is.na(STUDENTS)))
```

SEGUIR ejemplo clase pima.R buscando means y sds en otros R para saber que hacer en ese apartado